Confidence Score: A Data-Driven Measure for Inclusive Systematic Reviews Considering Unpublished Preprints

Authors:

Jiayi Tong, MS[1,12]
Chongliang Luo, PhD[2]
Yifei Sun, PhD[3]
Rui Duan, PhD[4]
M. Elle Saine, MD, PhD, MA[1]
Lifeng Lin, PhD[5]
Yifan Peng, PhD[6]
Yiwen Lu, MS[1,12]
Anchita Batra[1],
Anni Pan[1],
Olivia Wang[1],
Ruowang Li, PhD[1]
Arielle Anglin, PhD[1]
Yuchen Yang, PhD[1]
Xu Zuo, MS[7]
Yulun Liu, PhD[8]
Jiang Bian, PhD[9]
Stephen E. Kimmel, MD, MSCE[10]
Keith Hamilton, MD[1]
Adam Cuker, MD, MS[1]
Rebecca A. Hubbard, PhD[1]
Hua Xu, PhD[11,*]
Yong Chen, PhD[1, 12*]

Affiliation of the authors:

[1]Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA
[2]Division of Public Health Sciences, Washington University School of Medicine in St. Louis, St Louis, MO, USA
[3]Department of Biostatistics, Columbia University, New York City, NY, USA
[4]Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, MA, USA
[5]Department of Epidemiology and Biostatistics, University of Arizona
[6]Population Health Sciences, Weill Cornell Medical College
[7]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
[8]University of Texas Southwestern Medical Center, Dallas, TX, USA

[9]College of Medicine, Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, FL, USA
[10]College of Public Health & Health Professions and College of Medicine, University of Florida, Gainesville, FL, USA
[11]Yale School of Medicine, New Haven, CT, USA
[12]The Center for Health Analytics and Synthesis of Evidence (CHASE), The University of Pennsylvania, Philadelphia, PA, USA

Background

Since its emergence in December 2019, COVID-19 has had a global impact. As of June 16, 2023, there have been over 360,000 COVID-19-related manuscripts published on PubMed and preprint servers such as medRxiv and bioRxiv, with preprints making up approximately 15% of all manuscripts. Here, we define preprints as unreviewed manuscripts written in the style of a peer-reviewed journal article[1]. Compared to peer-reviewed articles, preprints can speed up dissemination by eliminating the months-long interval between submitting a manuscript to a publisher to the first public release of the manuscript. Preprints can also collect public comments to improve the rigor of the work, provide new opportunities to form new scientific collaborations, and avoid publication bias since preprints are issued at the discretion of the author. However, there is a lack of guidance on how to appropriately use results from preprints in systematic reviews[2].

In this study, we introduce, for the first time, a data-driven approach for assigning a "confidence score" to preprints in systematic reviews. The confidence score is intended to reflect the likelihood of a preprint surviving the peer review process. To determine the confidence score, we use "timestamps" for each preprint (i.e., the date it was posted and the date it was published, if applicable). By studying the "life cycle" of preprints in this way, we can examine factors such as the length of time a preprint is posted before being published and the impact of study-level and preprint-specific features (e.g., the number of patients in the study, citations of the preprint) on the likelihood of publication. We hypothesize that by appropriately modeling the life cycle of preprints using timestamps and relevant features, we can accurately predict the likelihood of a preprint being published. One benefit of the confidence score is that it is naturally defined on a scale from 0 to 1, which can be used as a weight in a subsequent meta-analysis.

Methods

We use a statistical model, known as a survival mixture model, to predict the probability that a preprint will eventually be published. Denote by T the time from posting to publication. For a preprint with features x, $\pi(x)$ represents its probability of being published, or in other words, the probability of T being finite given features x. This probability may be related to various features, including study-level features (e.g., number of patients, median age, RCT or observational study, single or multi-center study) and preprint-specific features (e.g., citations of the preprint during the first two weeks on a preprint server, number of downloads during first two weeks on preprint server).

We let $S_T(t|x)$ be the probability of $T > t$ for those that will be published, which depends on features x. The mixture model assumes that the probability that a preprint has not been published by time t is:

$$S_T(t|x) = \pi(x)S(t|x) + \big(1 - \pi(x)\big) \quad (1)$$

We use logistic regression with a logit link to model $\pi(x)$ and proportional hazards (PH) regression to model $S(t|x)$, i.e., $\pi(x) = \exp(x\gamma)/(1+\exp(x\gamma))$ and $S(t|x) = S_0(t)^{\exp(x\beta)}$, where β is the coefficient of the effects of x, and $S_0(t)$ is baseline survival function. Using model (1), we can predict the likelihood, defined as the confidence score, that a preprint will be published.

Using the confidence scores, we can synthesize both preprints and published articles in a meta-analysis. Suppose there are n total studies. We use w_i to denote the weight for the ith study in the meta-analysis, $i = 1, \ldots, n$. If the ith study is already published at the time of analysis, we set w_i to be 1. If the ith study is not published yet, its probability of never being published is estimated from the survival mixture model. The weight of the i-th study in the meta-analysis is set as $w_i = \{S(u_i|x_i) - 1 + \pi(z_i)\}/S(u_i|x_i)$ if it is not published. Using the *multiple imputation* procedure for evidence synthesis, we can obtain a final estimate for the overall effect size by taking the average of the estimated effect sizes for all imputations, as well as a final estimate for the heterogeneity variance.

Results

Study selection

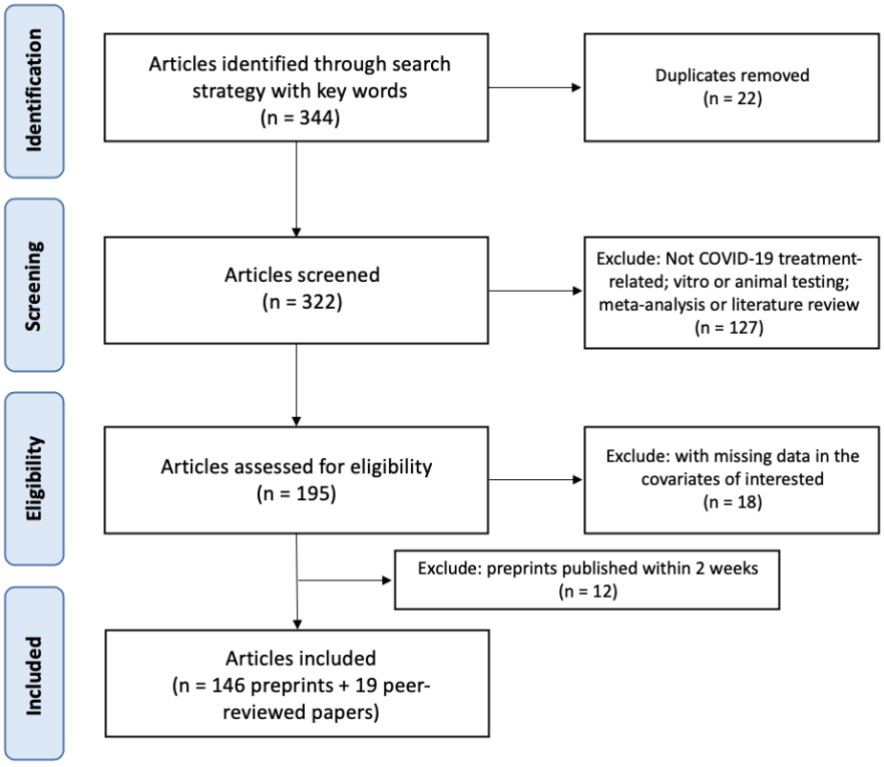

Figure 1. Literature review and study selection flow diagram.

We identified 146 preprints related to COVID-19 treatments, which were collected up to

04/30/2021. The literature review flow diagram is presented in **Figure 1**. To calculate the confidence score for the preprints, we extracted 11 variables with no missingness from the articles, including: whether or not the study was an RCT, median age of participants, the sample size, single or multi-center study design, whether or not the results were preliminary, whether or not the analysis was adjusted for confounding variables, country of study (with China, the US, Europe, and other countries as the reference group), whether or not the study was observational, h-index of the last author, citation counts, and PDF download counts in first two weeks posted on the server. Out of the 146 preprints, 84 have been published by 04/30/2021.

Confidence scores derived from metadata of 146 preprints

The estimated coefficients of the cure probability model and proportional hazards model with corresponding 95% CIs and p-values are presented in **Table 1**. In the sensitivity analysis, we excluded one variable at a time to calculate the confidence score. The results showed that citation counts had the greatest impact on the performance of the predictive model.

For example, the effect size of the “country -- China” variable is 0.53, which can be interpreted as that keeping all other variables constant, the confidence score of a preprint utilizing a study cohort from China is 1.70 (=exp(0.53)) times that of a preprint utilizing data from other countries other than the US and Europe or multiple countries. However, there exists evidence showing that this effect is not statistically significant with a p-value of 0.93 and a confidence interval covering zero. Across all the variables, the citation count is the only variable that is significantly associated with the confidence score with a p-value of 0.07. In the sensitivity analysis, we excluded one variable at a time to calculate the confidence score. The results showed that citation counts had the greatest impact on the performance of the predictive model.

Similarly, the effect sizes of the proportional hazards (PH) regression model are the estimated coefficients, denoted as β in the survival component in the survival mixture model, and are interpreted as the log hazard ratios. To illustrate, considering the "preliminary results" variable, where the hazard of being published of a preprint, which presents preliminary results, is about 1.36 (=exp(0.31)) times the hazard of another preprint that either doesn't present preliminary results or makes no mention of them, though there exists evidence showing that this effect is not statistically significant.

Table 1. Estimated effect sizes of the variables from the survival mixture model.

Variable	Effect size (95% CI) of mixture model	P-value	Effect size (95% CI) of PH model	P-value
RCT	4.44 (-14.42, 23.3)	0.64	-0.15 (-1.34, 1.03)	0.80
Median age	-0.17 (-0.41, 0.08)	0.18	0.04 (-0.02, 0.10)	0.15
Sample size	-0.37 (-1.82, 1.07)	0.61	0.19 (-0.26, 0.64)	0.41

Single center	0.45 (-5.77, 6.66)	0.89	0.39 (-0.79, 1.57)	0.51
Preliminary results	1.60 (-4.58, 7.77)	0.61	0.31 (-0.82, 1.44)	0.59
Adjusted analysis	1.48 (-5.35, 8.32)	0.67	-0.75 (-1.89, 0.39)	0.20
Observational study	2.64 (-8.12, 13.39)	0.63	-0.49 (-1.77, 0.79)	0.45
PDF downloaded counts	-0.18 (-2.33, 1.97)	0.87	-0.26 (-0.61, 0.10)	0.15
Country – China	0.53 (-11.76, 12.82)	0.93	-1.04 (-3.03, 0.95)	0.30
Country – US	1.50 (-9.6, 12.59)	0.79	-0.97 (-3.27, 1.32)	0.41
Country – Europe	3.13 (-5.55, 11.8)	0.48	-0.31 (-1.86, 1.23)	0.69
Last author h-index	1.97 (-1.83, 5.76)	0.31	-0.10 (-1.1, 0.89)	0.84
Citation counts	4.29 (0.48, 9.07)	0.07	0.11 (-0.05, 0.28)	0.18

Predictive performance of confidence scores

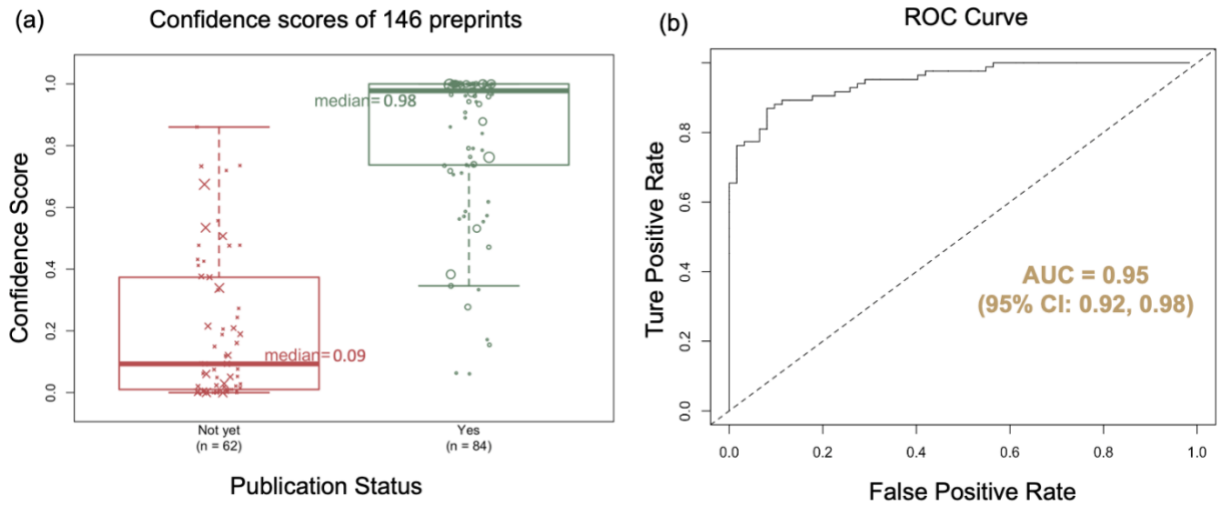

Figure 2. Predictive performance of confidence scores. (a) Box plot of estimated confidence score, stratified by publication status; (b) ROC curve of confidence score in predicting the publication status

Figure 2(a) shows the separation of confidence scores plotted against actual publication status. The median confidence score for unpublished preprints is 0.09, while for published preprints it is 0.98, which is significantly higher. **Figure 2(b)** presents the in-sample receiver operating characteristic (ROC) curve of the predictive model, with an area under the curve (AUC) of 0.95 (95% CI: 0.92, 0.98). Due to the limited number of total studies, we conducted leave-one-out cross-validation instead of splitting the data into training and test sets. The AUC value is 0.83 (95% CI: 0.75, 0.89). Overall, our pilot study has shown that the confidence score has the potential

to be a useful predictive measure of the likelihood of publication, even with a limited number of features extracted from preprints.

Conclusion

Our proposed confidence score has the potential to improve systematic reviews of evidence related to COVID-19 and other clinical conditions by providing a data-driven approach to including unpublished manuscripts. It is important to note that our method does not aim to replace existing measures of study quality but rather serves as a supplementary measure that overcomes some limitations of current approaches.

We acknowledge that the peer-review process can introduce modifications to a manuscript's content and structure. In the context of our method, we considered the fact that the fundamental findings and major conclusions presented in a preprint should remain consistent throughout the subsequent revisions for publication. While minor edits and refinements are expected during the peer-review process, our methodology is designed to focus on the core research outcomes and conclusions.

Regarding the applicable domains, our proposed approach is not limited to the COVID-19 domain. While we chose to apply it within the context of COVID-19 due to the significant "infodemic" surrounding COVID-19-related literature in recent years, the proposed method can be widely adaptable to other domains given the model flexibility in variable inclusion and its data-driven nature. In the future, when more features could be extracted and studies to be significant in a specific domain, then domain-specific lists of features could be built to construct the confidence scores for such domains and the model should be retrained for these specific domains.

We are dedicated to propelling the goals of OHDSI forward, continuously striving to improve and expand the proposed confidence score as cutting-edge data science methodologies for generating and synthesizing clinical evidence. Our commitment lies in refining and enhancing this powerful tool, which represent the future of evidence generation and synthesis in the field of healthcare.

References

1. NOT-OD-17-050: Reporting Preprints and Other Interim Research Products. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-050.html.
2. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, (2021).